Report on methodology and criteria followed for the selection of resources

Deliverable D2.4

Version 1.5

2011-11-30

Editor: Asunción Moreno

METANET4U

www.metanet4u.eu

The central objective of the Metanet4u project is to contribute to the establishment of a pan-European digital platform that makes available language resources and services, encompassing both datasets and software tools, for speech and language processing, and supports a new generation of exchange facilities for them.
This central objective is articulated in terms of the following main goals:
Assessment: to collect, organize and disseminate information that permits an updated insight into the current status and the potential of language related activities, for each of the national and/or language communities represented in the project. This includes organizing and providing a description of: language usage and its economic dimensions; language technologies and resources, products and services; main actors in different areas, including research, industry, government and society in general; public policies and programs; prevailing standards and practices; current level of development, main drivers and roadblocks; etc.
Collection: to assemble and prepare language resources for distribution. This includes collecting languages resources; documenting these language resources; upgrading them to agreed standards and guidelines; linking and cross-lingual aligning them where appropriate.
Distribution: to distribute the assembled language resources through exchange facilities that can be used by language researchers, developers and professionals. This includes collaboration with other projects and, where useful, with other relevant multi-national forums or activities. It also includes helping to build and operate broad inter-connected repositories and exchange facilities.
Dissemination: to mobilize national and regional actors, public bodies and funding agencies by raising awareness with respect to the activities and results of the project, in particular, and of the whole area of language resources and technology, in general.

METANET4U is a project in the META-NET Network of Excellence, a cluster of projects aiming at fostering the mission of META. META is the Multilingual Europe Technology Alliance, dedicated to building the technological foundations of a multilingual European information society.

Deliverable 2.4: Report on methodology and criteria followed for the selection of resources

METANET4U is co-funded by the participating institutions and the ICT Policy Support Programme of the European Commission

and by the participating institutions:

Faculty of Sciences, University of Lisbon

Instituto Superior Técnico

University of Manchester

University *Alexandru Ioan Cuza*

Research Institute for Artificial Intelligence, Romanian Academy

University of Malta

Technical University of Catalonia

Universitat Pompeu Fabra

Revision History

Version	Date	Author	Organisation	Description
V1.0	Oct.10, 2011	A. Moreno	UPC	First draft
V1.1	Nov 11, 2011	A. Moreno	UPC	Updated partner contributions
V1.2	Nov 16, 2011	A. Moreno	UPC	Updated partner contributions
V1.3	Nov 28, 2011	A. Moreno	UPC	Guidelines improved
V1.4	Nov 30, 2011	A. Moreno	UPC	Pre final to review QC
V1.5	Nov 30, 2011	A. Moreno	UPC	Final

METANET4U

Report on methodology and criteria followed for the selection of resources

Document METANET4U-2011-D2.4
EC CIP project #270893

Deliverable
Number: D2.4
Completion: Final
Status: Submitted
Dissemination level: Restricted to project participants

Responsible: Asunción Moreno (WP2 coordinator)

Contributing Partners: FCUL, IST, UNIMAN, UAIC, RACAI, UOM, UPC, UPF

Authors: António Branco, Amália Mendes, Isabel Trancoso, Hugo Meinedo, Sophia Ananiadou, Paul Thompson, John McNaught, Dan Cristea, Diana Trandabat, Dan Tufis, Mike Rosner, Asunción Moreno, Núria Bel, Jorge Vivaldi, Eva Revilla

Reviewer: Paul Thompson

Contents

1 Introduction

The central objective of the Metanet4u project is to contribute to the establishment of a pan-European digital platform that makes available language resources and services, encompassing both datasets and software tools, for speech and language processing, and supports a new generation of exchange facilities for them. This central objective is articulated in terms of several main goals. One of these main goals is the collection of language resources, i.e., the assembly and preparation of language resources for distribution. This goal includes gathering language resources — both endogenous, that are owned or controlled by the project partners, and exogenous, that are not directly managed or controlled by them —; documenting these resources; upgrading them according to agreed standards and guidelines; and linking and aligning them cross-lingually where appropriate.

WP2 and WP3 of the Metanet4u project are directly concerned with the goal of collecting language resources. WP2 takes care of the analysis and selection of language resources, while WP3 deals with the enhancement of language resources. The objective of this *Deliverable D2.4 Report on methodology and criteria followed for the selection of resources* is to describe the overall guidelines that have been followed in the selection of these language resources.

This deliverable is organized as follows: section 2 explains the general guidelines followed for the selection of language resources, while section 3 shows, for each partner, the resources that have been selected by following these criteria.

2 Guidelines for the selection of resources

Consortium composition: Languages involved

There are five national languages represented in this project, i.e., English, Maltese, Portuguese, Romanian and Spanish, and other co-official languages of Spain. This permits three different language families to be addressed, i.e., Germanic, Romance and Semitic, and allows for an analysis of significant issues raised by language diversity. It is also important to note that the set of languages covered includes one of the most studied and, in terms of language resources, one of the best equipped languages in the world, i.e., English. In contrast, one of the languages with the least resources is also represented, i.e., Maltese. This will permit relevant methodological issues to emerge, and fast developmental techniques to be exercised.

The consortium consists of a set of universities and research centres that are very well known both nationally and internationally. At the national and regional levels, there is a widely shared perception of language as a key element of membership (cultural, national, etc.) and as a crucial factor of identity to be preserved, particularly in these times of rapid change and globalization. This has led to a growing number of national and regional initiatives to work towards more focussed language policies, which include specific programmes both for research into language science and technology and for the

deployment of language resources. An important issue regarding the selection of language resources is the detailed knowledge that the consortium partners possess about other national and regional projects, initiatives, actors and agencies.

Among the languages represented in the project, there are the three European languages, i.e., English, Portuguese and Spanish, that having the largest number of worldwide native speakers, the majority of whom are outside the EU. Building on the large network of overseas professional relations of the consortium partners, this circumstance is being explored as a strategic vantage point for the project. This will permit the number of language resources attracted to the project to be enlarged, and will also allow resources that deal with non-European variants of these languages to be represented. This, in turn, will permit the range of activities and target groups for the dissemination of the project results to be widened, and for their outreach to be amplified. It should be noted that all languages represented in the Metanet4u project are equally important, regardless of the number of speakers.

Attracting exogenous resources: Networking

In addition to the resources already known by the partners, several strategies to attract exogenous resources to be distributed through the META-Share exchange platform will be adapted to the specific to the circumstances of each national and language community. Particular approaches for this adaptation are outlined below.

English (European and non-European variants): An important remit of the UK National Centre for Text Mining (NaCTeM), at the University of Manchester, is to act as a hub for text mining research and services, both in the UK and abroad. It provides text mining services, resources and tools, not only to the UK academic community, but also to the international community in a multitude of domains, such as biology, medicine, chemistry, education, media and social sciences. As a national hub, they are in excellent position to attract further resources, thus making them visible to the wider community with the aid of the exchange platform of the META-NET network.

Maltese: Although Malta is small, it is characterised by a large number of institutions that are potential sources of written and spoken language resources that work independently of each other. Our general strategy will be (i) to identify all the stakeholders; (ii) to hold one or more meetings whose goals are to convey the aims and objectives of the project; (iii) to persuade stakeholders of the validity of a common goal; and (iv) to develop a local language resource portal for the submission, management, and annotation of resources coming from different sources.

Portuguese (European and non-European variants): For the Portuguese language, the networking aimed at attracting resources to the upcoming META-NET distribution platform will be undertaken by the two teams at the University of Lisbon, i.e., the Center of Linguistics and from the Department of Informatics, and by the third team at IST. The networking will have an international dimension and will seek to attract resources from both the American and the European variants of Portuguese ,when this difference may be relevant for the type of resource at stake. This international networking will build on the strong professional links established in the scope of the community working on the

computational processing of Portuguese, the Portuguese national CLARIN network and the IST team connections.

Romanian: UAIC and RACAI established a list of researchers working in the Romanian language technology and resources, who can be approached to help them release their resources through META-NET. This list includes the linguists and computer scientists grouped within the two major professional organisations in Romania concerned with creating and/or using language resources and tools for Romanian: The Commission of the Romanian Academy for Language Technologies for Romanian, and its executive body, called the Consortium for Resources and Technologies for Romanian (ConsILR), and The Romanian Terminology Association (TERMROM). In addition, we plan a wide investigation of the available language resources and tools for Romanian that exist both in Romania and abroad.

Spanish (European and non-European variants) and other co-official languages: We are going to approach the Spanish Thematic Network on Speech Technologies (acronym RTTH in Spanish). RTTH is a network funded by the Spanish Government and since 1999, more than 250 researchers from more than 40 research groups from Spanish universities and companies have joined the network. Importantly, this can also act as the seed for the integration of resources in other languages (Catalan, Basque and Galician). There are other networking initiatives that will facilitate the attraction of resources to the project in Spain, namely the MAVIR Consortium, and the TIMM network. CLARIN-ES has been the latest initiative for Language Resources and Technologies Networking, with a specific interest in the European and international scenario. CLARIN-ES networking now has about 16 institutional members. A network of Latin American universities from Venezuela, Colombia, Costa Rica, Chile and Argentina was created in the framework of data collection SALA projects. We plan to interact with this international network to get access to new collections of resources.

Technology enhancement

While the starting point for the work to be undertaken in this project is the existing prototype services and language resources, either endogenous or exogenous, which will be made available through new generation dissemination channels, a key endeavour to be pursued is also to leverage to a new level the potential of these resources to support language-technology based products and services. This will be achieved by establishing links across different types of resources and modalities and by linking a core set of resources across different languages. The core set of resources selected in this project over which such cross-linking will be established involve a range of modalities, e.g., text, speech and video; linguistic dimensions, e.g., phonology, morphology, syntax and semantics; data set types, e.g.,
corpora, lexica and treebanks; and basic processing tasks, e.g., speech recognition, word sense disambiguation and parsing. They include aligned lexical semantic networks, parallel treebanks and multimodal corpora. For the languages represented in this project, these cross-linked resources will contribute to an improved capacity for language technology to be at the heart of global online services, which are able to overcome language barriers. By way of an example, immediate multilingual applications that can be supported by these resources include machine translation or cross-lingual

information retrieval systems. However, it is important to note that the classes of resources in question, i.e., aligned lexica, parallel and multimodal corpora, lie at the heart of virtually any speech and language processing task and tool, ranging from speech recognition to deep linguistic processing, and including morphological analysis, named-entity recognition, syntactic parsing, word sense disambiguation or semantic role labelling, among many others. These tools provide the building blocks for major language technology applications such as automatic subtitling, speech synthesis, speaker detection, terminology extraction, sentiment analysis, text clustering, summarization, question answering, natural language interaction, etc., besides the aforementioned machine translation and information extraction. By cross-linking the core set of resources selected in this project, their potential will be lifted to a new level so that, when combined with further appropriate technologies, they will form one of the key elements underpinning products and services operating seamlessly in the web, which in its content is heavily language based and multilingual. Naturally, this will have a direct impact on the translation and localization sector, with obvious economies of scale resulting for the increased quality of the underlying tools. However, even though less visible for the final user or consumer, it concerns also other socio-economic sectors that are either required to operate in a multilingual context (from the audiovisual industry to public services) or that are able to explore the advantages of reaching a multilingual target audience (from the telecommunications sector to the advertising business).

Expected audience

At the head of the expected audience of the project results will be the group of technology developers who are engaged in the fostering of innovation processes. For these users, the language technology solutions supported by the resources released by the project offer an opportunity to leverage the operation of their services and products, in to reach new customers in a online multilingual environment, to explore innovative business opportunities, to obtain economies of scale, etc. The language resources and tools made available will also be of utmost interest to the group of researchers and academics for whom the language resources to be released are important instruments for the advancement of their research and development activities. This group includes linguists and other scientists of language engaged in the study of the structure and function of natural languages, as well as scientists from other academic disciplines where natural language plays a crucial role. This group additionally includes language technologists working on the computational processing of natural languages and on the data sets and tools that help to improve such processing. Finally, the materials being released may be also of interest of language professionals such as localization professionals, translators, etc. For some high level, specialized tasks and problems, these professionals may find that the language resources, data sets and software released by the project can provide assistance in their everyday activities.

Language Resources and Technologies

The term *Language Resources* encompasses not only data resources but also software and tools, i.e., the tools, technologies, and services used for processing datasets. The term is also found in the literature as Language Resources and Technologies (LRTs). Within this deliverable, we differentiate the criteria and selection of the language

resources, depending on whether they belong to the data group or to the software and tools group:

Data Resources
This group includes among others: lexica, wordnets, thesauri, annotated corpora, parallel corpora, multimodal and multimedia data, grammars, language models, speech recognition databases and speech synthesis databases.
The general criteria used to select data resources for a given language can be summarised as follows:

- for a range of typical types of datasets (e.g. lexica, corpora, etc.), try to include datasets such that every such type is represented;
- for a range of typical levels of annotation or metadata (e.g. POS, treebanking, propbanking, etc.), try to include datasets such that every such level type is represented;

Software and tools
This group includes, among others, language identifiers, hyphenizers, tokenizers, lemmatizers, sentence splitters, POS taggers, NP-chunkers, parsers, semantic role labellers, summarisers, word aligners, lexicon editors, linguistic web services, workflow platforms, grapheme to phoneme converters, etc.
The general criteria used to select software and tools for a given language can be summarized as follows:

- for a whole pipeline of processing (e.g. tokenization to deep linguistic processing, etc.), try to include at least one software tool/service for each significant intermediate step;
- for a range of typical software applications (e.g. summarization, machine translation, etc.), try to include at least one software tool/system for each type of application;

The above general guidelines were evaluated against the actual situation in terms of existing resources for each language represented in the project. They were adjusted to take into account the current availability and quality of resources, and the perceived potential for reuse, recombination and repurposing of the resources.

The guidelines were also adjusted to take into account the specific goals of WP3 concerning documentation, and possible upgrading or extension according to agreed standards and guidelines.

3 Selection of resources by partners

This section lists the language resources selected by partners according to the criteria outlined above. There are approximately 150 endogenous resources and over 70 exogenous resources. As the latter resources are not owned or controlled by the project partners, their inclusion in the project collection of resources will be dependent on third parties. In order to attract these resources to the project activities under terms that are acceptable to their owners, they will be approached according the guidelines described

in the previous section. This set of resources is further complemented with unrestricted exogenous resources, which are not owned by the project partners, but are freely available. We have identified 20 such resources, which represent a first step in listing unrestricted exogenous resources.

The selected language resources listed in this section are grouped by partner. The resources are divided into two groups, i.e. *Data resources* and *Software and tools*.

In order to clarify the link between the selected datasets and the selection criteria that they fulfil, the list groups the data resources according to their dataset type and subtype. Following the name of each resource, a short description of the annotation level is also provided. The layout of the information about resources is as follows:

Resource type

Sub-type

Resource name

Annotation level

Software and tools are divided into *LT Tools* and *Services*; the list includes the name of each tool/resource, together with a short description

3.1. Partner: ULX

3.1.1 Data resources

annotated corpus

Portuguese corpus

- CETEMPúblico
  - POS
- CINTIL-Internacional Corpus of Portuguese
  - orthographic form, POS, nominal inflection, verbal inflection, lemma, IOB named-entity
- COMPARA
  - POS
- Corpus NILC
  - Syntactic annotation
- CorpusTCC
  - RST annotation
- DEF Corpus
  - POS, lemmas, semantic tag, keywords, definitions
- DependencyBank
  - form, lemma, POS, inflection, grammatical functions
- PAROLE corpus
  - metadata and POS annotation of a subset of 250.000 tokens following encoding standards established by the PAROLE consortium
- PLN-BR Gold
  - POS, semantic annotation
- PropBank
  - Penn PropBank tagset
- RHETALHO
  - RST annotation
- Summ-it
  - POS, correference
- Treebank
  - POS, syntactic constituency

dictionary

medieval

- Dicionário de Verbos do Português Medieval (DVPM)
  - No annotations

grammar

training models for tagger

- NILC Taggers
  - No annotations

lexicon

frequency lexicon

- Multifunctional Computational Lexicon of Contemporary Portuguese - CORLEX
  - POS, lemma, with quantitative information (information for frequency levels and frequency exact values)

lexical database

- Glossário
  - No annotations
- MorDebe
  - orthography, verbal inflation and morphological relations between words

lexical database with information on word structure

- PORLEX
  - number, orthographic wordform, diacritic pointer, phonetic wordform, variant pointer, grammatical class, open/close class, stress position,

letter length, orthographic syllable length, phonemic length, hormorganic nasals, phonological syllable length, orthographic syllabication, phonological syllabication, gender, inflexion, lexical frequency, and others.

lexicon

PAROLE lexicon

syntactic and morphological annotation, following standards established by the PAROLE consortium

SIMPLE lexicon

concept type classification, distribution by class and semantic domains, synonym relations, predicative structure definition, argument structure and semantic roles constraints; each semantic unit is linked to its syntactic (and morphological) encoding

list of abbreviations

Abbreviations

No annotations

list of stopwords

Stopwords

No annotations

ontology

geographic terms

Geo-Net-PT01

No annotations

scientific terms

Ontologia de Nanociência e Nanotecnologia

No annotations

raw text corpus

parallel

European Parliament Parallel Corpus

No annotations

The JRC-Acquis Multilingual Parallel Corpus

No annotations

text

Clássicos LP/Porto Editora

No annotations

Corpus NILC

No annotations

Technical Corpus

Metadata

TeMário 2006

human summaries

speech

spoken **corpus**

C-ORAL-ROM Portuguese Corpus

POS, Lemma, terminal and non-terminal prosodic breaks, metadata, text-to-sound alignment at the utterance level

Norma Urbana Culta (NURC)

No annotations

Panorama do Português Oral de Maputo (PPOM)

No annotations

PF Corpus

Metadata

Spoken Portuguese

text-to-sound alignment, metadata

wordnet

wordnet

MWN.PT

n.a.

3.1.2 Software and LR tools

LT tool

discourse parser
DiZer 2.0
ontology buidling
Ontolp Plugin
POS tagger
Tagger
sentence and word aligners
Text Aligners
sentence splitter
Chunker
stemmer
Stemmer
summarizer
GistSumm
tagger
Forma
tokenizer
Tokenizer

3.2. Partner: IST

3.2.1 Data resources

annotated corpus

multilingual word alignments
PTSTAR golden collection
Language-pairs annotations created with the annotation and visualization tool implemented by Chris Callison-Burch (University of Edinburgh)

raw text corpus

text
Named entities tagged in natural language questions
Category XML tags

speech

speech database
CORAL
A subset of the corpus has been annotated in several levels (orthographic, phonetic, phonological, syntactic and semantic). Orthographic transcription has been done for the whole corpus.
LECTRA
A subset manually transcribed with multi-layer annotations

3.3. Partner: UNIMAN

3.3.1 Data resources

annotated corpus

event annotation
GREC
Events whose arguments are based on 13 semantic roles, i.e. theme, agent, manner, destination, source, location, rate, condition, rate, purpose, descriptive-theme, temporal, instrument, descriptive-agent

GENIA event corpus
Biomedical event annotation enriched with meta-knowledge annotation which determines how each event is to be interpreted

POS annotation and terms
GENIA
1999 MEDLINE abstracts with POS and biomedical term annotation. POS annotation generally follows the Penn Treebank POS tagging scheme with some modifications

lexicon

large-scale terminological resource
BioLexicon
semantic types: proteins, genes, enzymes, protein domains, chemicals, diseases, molecular roles, cell, operons, sequences. Semantic relations, variants, synonymy relations. verbal syntactic subcategorisation frames, verbal semantic event frames .

verbs
SemLink Resources
mapping between the syntactic/semantic subcat frames of VerbNet, PropBank, FrameNet and WordNet,

3.3.2 Software and LR tools

LR tools

Deep Parser
U-Compare Enju Parser

named entity recogniser
NEMINE

PCFG Parser
U-Compare Stanford Parser

POS tagger
U-Compare GENIA PoS Tagger
U-Compare OpenNLP PoSTagger
U-Compare STEPP PoS Tagger

sentence splitter
U-Compare GENIA Sentence Detector
U-Compare NaCTeM Sentence Detector
U-Compare OpenNLP Sentence Detector

tagger, chunker and NER for biomedical text
Genia tagger/chunker and NER

tokenizer
U-Compare GENIA Tokenizer
U-Compare OpenNLP Tokenizer

Workflow management tool
U-Compare Workbench

Services

Specification of linguistic annotations
U-Compare Type System

3.4. Partner: UAIC

3.4.1 Data resources

annotated corpus

annotated corpus of noun phrases

1984_NP
NP and their head

annotated Question Answering corpus
QA-corpus-UAIC
question type, answer type, question focus, question keywords

correference annotated corpus
1984AnaphoraRo
co-reference chains among NPs

parallel French-Romanian multi-word expressions corpus
FrRoMWE
multi-word expressions and phraseological units annotated in French and Romanian

parallel English-Romanian semantic roles corpus
RO-FN
semantic roles annotated in English and Romanian, in FrameNet style

syntactic annotated corpus
RO-FDGBank
syntactic dependency trees

dictionary

Dictionary of adult education
DEA
lemma, part of speech, frequencies

Dictionary of poetic language in Eminescu's work
DLPE
lemma, part of speech, frequencies

grammars

collection of rules for classification of textual inferences
TE-rules
rules for determining the textual inference (yes, no, unknown)

collection of Text-Hypothesis pairs
TE-pairsResource-UAIC
pairs of texts and hypotheses annotated with textual inference (yes, no, uncertain)

paradigmatic morphology rules in symbolic form
RomMorph-UAIC
not applicable

lexicon

lexicography
eDTLR
dictionary entries, semantic fields, definitions, citations, etc.

lexicon
RoSemClass
words annotated with their semantic classes, from a fixed list of 30 classes

ontology

semantic dictionary
RoWN-eDTLR
linking between RoWN synsets and eDTLR senses

raw text corpus

collection of scanned and OCR-ed books
eDTLR-sources
none

speech

speech corpus: annotated and documented speech resource
SRoL – Sounds of the Romanian Language
orthographic transcription of really pronounced; noise marks (speaker noise, background noise, etc.): PRAAT style

3.4.2 Software and LR tools

LR tools

Diacritics Recovery System
Diacritics-UAIC
Discourse Parser system
DP-UAIC
Document category/domain identification
Categorizer-UAIC
Functional Dependency Parser
FDGparser-UAIC
Language identifier
Language identifier-UAIC
Lemmatizer
Lemmatizer-UAIC
Named Entities Recognizer
ANNIE
NLP workflow builder
ALPE-UAIC
NP-chunker
NP-chunker-UAIC
Occurrence Finder
Occurrence Finder-UAIC
Ontology Builder
OntologyBuilder-UAIC
Question Answering
QA-UAIC
Robust rule-based Anaphora Resolution system
RARE-RO-UAIC
Semantic Role Labeling
SRL-UAIC
sentence splitter
Splitter-UAIC
Summarization system
Summarizer-UAIC
Textual Entailment
TE-UAIC
tokenizer
Tokenizer-UAIC
Word flexing system
AnaMorph-UAIC

3.5. Partner: RACAI

3.5.1 Data resources

annotated corpus

comparable corpora
Multilingual News Corpus
RO, EN will be XCES compliant and fully marked up (lemma, tag, MSD, linkage)
largest annotated corpus for Romanian
Romanian Balanced Corpus
xml (XCES compliant)
parallel corpus
RO-SemCor

pos, MSD, lemma, WN sense number

Romanian Corpus

RO-Acquis

pos, MSD, lemma, delendency linking

subjectivity analysis

Multilingual Subjectivity Analysis: Gold Standard and Training Data

sentences are manually classified with Opinion Finder tool

written corpus

TimeBank parallel corpus

TimeML compliant

dictionary

Romanian reference explanatory dictionary

WEB-DEX

Concede encoding schema

grammars

n-grams

WEB 1T 5-gram

1-, 2-, 3-, 4- and 5-grams in 10 European languages including Romanian

lexical ontology

Semantic dictionary

RO-WordNet

XML

lexicon

Wordform lists

Wordform lexicons

wordform, lemma, MSD, POS

Indiosyncrasic items

CoDII-NPI.ro

syntactic information about the item itself and the licensing environment in which it occurs

speech

speech database

RO-SAM EUROM

Orthographic and phonetic transcriptions. xml CES

3.5.2 Software and LR tools

Software

audio segmentation and speech synthesis

VoiceForge

concordancer

Lucon

LR tools

a dependency parser for several languages (including Romanian) texts

VISL Dependency-Parser

chunker

TTL-chunker

collocation extractor

COLLOC

Dependency linker

LexPar

diacritics restorations for Romanian texts

DIAC+

hyphenator

RO-HYPHEN

language identification

LangId

Lemmatizer

TTL-lemmatizer
Lexical Chain
LexChain
Morpho-syntactic tagger
TTL-Tagger
tokenizer
TTL-Tokenizer
word aligner
YAWA
Word Sense Disambiguation
SynWSD
wordnet editor
WN-Builder

3.6. Partner: UOM

3.6.1 Data resources

dictionary

dictionary
Basic English-Maltese Dictionary
TEI-compliant XML
Busuttil Dictionary EN/MT
TEI-compliant XML
Busuttil Dictionary MT/EN
TEI-compliant XML
Malta Online Dictionary
TEI-compliant XML

lexicon

lexicon
Acquilina Dictionary MT/EN
initially none but should be developed
lexicon/Database
Eurowordnet
initially none but should be developed
lexicon/Knowledge Source
Maltese Wordlist
none
bilingual Lexicon
Combined Maltese/English Bilingual Lexicon
lex

raw text corpus

Maltese Company Data
MFSA_Companies_Register
n.a.
written
Illum_Corpus
none
Maltese Fiction
initially none but should be developed
written corpus
Laws of Malta
raw
Maltese Acquis Communautaire EN
XML
Maltese Acquis Communautaire MT

XML
MLRS Corpus
text structure / XML

speech

annotated speech corpus
Maltese Speech Engine Corpus
adequate for training of speech synthesiser

speech data/ spoken corpus
F_MONA_1 / Maltese Spoken Newspaper
phoneme duration(being built)

wordnet

WordNet
MaltiWordNet of Maltese – Preliminary version – 15,000 entries
ann

3.6.2 Software and LR tools

Services

Web Service
MLRS API
MLRS1 Corpus Manager
MLRS1 Lexicon Editor

LT Tools

POS Tagger
MLRS API - POS Tagger

3.7. Partner: UPC

3.7.1 Data resources

lexicon

ASR and TTS lexicon
LC-STAR CATALAN
Tagset: POS defined in LC-STAR, applicable to many languages.
LC-STAR SPANISH
Tagset: POS defined in LC-STAR, applicable to many languages.

raw text corpus

newspaper
EL_PERIODICO_97-07
Bilingual aligned texts

speech

multimodal speech database
CHIL UPC Interactive Seminars
Orthographic transcription of really pronounced speech, acoustic events type and endpoints, identity of speakers, position of persons.

speech database
AGORA
Verbatim orthographic transcription
ALBAYZIN
Orthographic transcription of really pronounced
BN RadioBCN
Verbatim orthographic transcription
Catalan-SpeechDat

Orthographic transcription of really pronounced; noise marks (speaker noise, background noise, etc.)

CatalanBN

Verbatim orthographic transcription

CHIL UPC Seminars

Orthographic transcription of really pronounced

EUROM.1

Orthographic transcription of really pronounced

FESTCAT

Orthographic transcription; Phonetic transcription: manual; Phonetic segmentations: automatic; Basic prosodic manual; Pitch labelling: automatic

FESTCAT-SEL

Orthographic transcription; Phonetic transcription: automatic; Phonetic segmentations: automatic; Basic prosodic none; Pitch labelling: automatic

FREE-SPEECH

Orthographic transcription

INTERFACE

Orthographic transcription

LAS CORTES

Verbatim orthographic transcription

LC-STAR Dialogues

Orthographic transcription of really pronounced

SALA-Mexico

Orthographic transcription of really pronounced; noise marks (speaker noise, background noise, etc.)

SALA-Venezuela

Orthographic transcription of really pronounced; noise marks (speaker noise, background noise, etc.)

SPANISH EPPS

Verbatim orthographic transcription

SpeechDat-Car Catalan

Orthographic transcription of really pronounced; noise marks (speaker noise, background noise, etc.)

SpeechDat-Car Spain

Orthographic transcription of really pronounced; noise marks (speaker noise, background noise, etc.)

Speecon Catalan

Orthographic transcription of really pronounced; noise marks (speaker noise, background noise, etc.)

Speecon Spanish (SVOX)

Orthographic transcription of really pronounced; noise marks (speaker noise, background noise, etc.)

TALP TTS0 BASELINES

Orthographic transcription; Phonetic transcription: manual female spk; Phonetic segmentations: manual female spk; Pitch labelling: manual female spk.

TC-STAR TTS BASELINES

Orthographic transcription; phonetic transcription: 20% of it manual; Phonetic segmentations: 20% of it manual; Basic prosodic 20% of it manual; Pitch labelling: 20% of it manual

TC-STAR TTS Expressive

Orthographic transcription; automatic phonetic transcription

TC-STAR VC

Orthographic transcription; automatic phonetic transcription

Spanish SpeechDat (M) and SpeechDat (II)

Orthographic transcription of really pronounced; noise marks (speaker noise, background noise, etc.)

synthetic speech

Text to speech synthesis
- Bilingual Speech synthesis
  - HTS models Festival compliant
- Spanish Festival models
  - HTS models Festival compliant
- Spanish Festival voices
  - Orthographic, Phonetic, Pitch, Festival compliant

3.7.2 Software and LR tools

LR tools

Multichannel Speech Recording Plataform
- NannyRecord

Rule-based phonetic transcription
- Saga

Synthetic waveform generation
- HTS

Text to speech synthesis system
- Festival

Services

Platform to integrate speech-to-speech translation components
- Gaia

3.8. Partner: UPF

3.8.1 Data resources

annotated corpus

Written + oral corpus
- Electronic Corpus of Academic Materials – University of Zaragoza (ECAM-UZ)
  - morphosyntactic

Written annotated corpus (POS-tagged)
- Corpus Técnico do Galego
  - morphosyntactic (EAGLES)
- Corpus CLUVI
  - structural (TMX), morphosyntactic
- Corpus PAAU 92
  - POS
- Genoma corpus
  - POS
- IULA Technical Corpus
  - POS-tagged
- SenSem Corpus
  - POS

Written annotated corpus (other tagged)
- AnCora-Ca
  - + coreference
- AnCora-Co-CA
  - morphological (PoS), syntactic (constituents and functions) and semantic (argument structure and thematic roles, semantic class, named entities and WordNet senses)
- AnCora-Co-ES
  - morphological (PoS), syntactic (constituents and functions) and semantic (argument structure and thematic roles, semantic class, named entities and WordNet senses)

AnCora-Es
morphological (PoS), syntactic (constituents and functions) and semantic (argument structure and thematic roles, semantic class, named entities and WordNet senses)
All resulting layers are independent of each other
CESS_EU: The Basque Dependency Treebank
Dependency annotation
Computer Science Tri-lingual Corpus
annotated morphologically and syntactically

lexica

Lexical Resource - equivalents

Basic Vocabulary on the Human Genome
not annotated
Multilingual Vocabulary of Economics
not annotated

Lexical resource - neologisms

Neologisms of the year: Bank of Spanish and Catalan Neologisms
not annotated

lexicon

Apertium Basque dictionary
POS
Apertium Bilingual dictionary , Basque-Spanish,
Bilingual correspondences
Apertium Bilingual dictionary CA-ES
Bilingual correspondences
Apertium Bilingual dictionary English-Catalan
Bilingual correspondences
Apertium Bilingual dictionary English-Galician
Bilingual correspondences
Apertium Bilingual dictionary English-Spanish
Bilingual correspondences
Apertium Bilingual dictionary French-Catalan
Bilingual correspondences
Apertium Bilingual dictionary French-Spanish
Bilingual correspondences
Apertium Bilingual dictionary Occitan-Catalan
Bilingual correspondences
Apertium Bilingual dictionary Occitan-Spanish
Bilingual correspondences
Apertium Bilingual dictionary Portuguese-Catalan
Bilingual correspondences
Apertium Bilingual dictionary Portuguese-Galician
Bilingual correspondences
Apertium Bilingual dictionary Spanish-Asturian
Bilingual correspondences
Apertium Bilingual dictionary Spanish-Galician
Bilingual correspondences
Apertium Bilingual dictionary Spanish-Portuguese
Bilingual correspondences
Apertium Bilingual dictionary Spanish-Romanian
Bilingual correspondences
Apertium Catalan dictionary
POS
Apertium Galician dictionary
POS
Apertium Spanish dictionary
POS
Dicionario CLUVI inglés-galego
Bilingual correspondences

Euskal Wordnet 3.0
POS; language-internal relations to other synsets
FreeLing Asturian dictionary
POS
FreeLing Catalan dictionary
POS
FreeLing Catalan sense dictionary
WordNet-based sense dictionary, extracted from EuroWordNet (reduced subset)
FreeLing Galician dictionary
POS
FreeLing Spanish dictionary
POS
FreeLing Spanish sense dictionary
WordNet-based sense dictionary, extracted from EuroWordNet (reduced subset)
PAROLE lexicon
syntactic and morphological annotation, following standards established by the PAROLE consortium
SenSem Database
sense dictionary
SIMPLE lexicon
concept type classification, distribution by class and semantic domains, synonym relations, predicative structure definition, argument structure and semantic roles constraints; each semantic unit is linked to its syntactic (and morphological) encoding
Spanish FrameNet
sense dictionary
Spanish Wordnet 3.0
WordNet-based sense dictionary

Terminological Resource

European Community Law Catalan Glossary mapped to EUROVOC
not annotated
Termoteca
not annotated

Terminology bank

UPF_Term
not annotated

parallel corpus

Written aligned

Parallel IULA Technical Corpus
POS-tagged

Aligned Corpus

DOGC CAT-SPA Parallellized Corpus
Parallelized corpus

Questions for Question Answering Classification

6305-QC
Manually tagged with the Hoby categories

3.8.2 Software and LR tools

LR infrastructure tools

Cleaning and pre-processing

Tools for automatic UTF-8 conversion, stand-off and LMF conversion

LR tools

Corpus Processing

Tools for Catalan Corpus Processing
Tools for Spanish Corpus Processing